How Do the Engineer Students Learn the SQL Language?

M. Czenky
Szent István University, Gödöllő, Hungary

Abstract—**In the teaching of database management the teaching of standardized SQL language cannot be avoided. Until the appearance of e-learning education systems we could only support the learning of SQL language with showing example queries. Using e-learning education systems in the education enables using tests, tutorials, teacher feedbacks which facilitate the learning of SQL language. But still these education tools can not ensure that students receive error message and assessment for correctness of the query edited in database management system in their native language after running the SQL query. In order to accomplish these two latter purposes we wrote a SQL tutoring system which was used from 2012's autumn by the students. In the paper we describe the problems of learning the SQL language, the supporting methods for learning before installation of the SQL tutoring system and with analysis of data of log file of the SQL tutoring system the manner of students' learning of SQL.**

Index Terms—**SQL, e-learning, tutoring system, survey, examination, assessment**

I. INTRODUCTION

At the Faculty of Mechanical Engineering and Faculty of Agricultural and Environmental Sciences of Szent István University we teach database management for mechanical engineer, engineering manager and environmental engineer students in both BSc and MSc qualification.

We summarized the characteristics of taught subjects in Table I.

TABLE I.
THE CHARACTERISTICS OF TAUGHT SUBJECTS

Name of subject	Abbreviation	Major	Qualification/ Course	Class per week lecture+ practise	Year
Applied Informatics	MM	engineering manager	MSc, correspondence	0+2	2010-2012
Computer Studies III.	KM3	environmental engineer	BSc, full time	2+2	2010-2012
Database Management	ABK	mechanical engineer	BSc, full time	0+2	2007-2010
Environmental Databases	KDB	environmental engineer	MSc, full time	2+2	2011-2012

II. THE JUDGMENT OF THE LEARNING OF SQL LANGUAGE

SQL language is highly structured and has limited number of statements. An SQL statement is much shorter than the programs written in some procedural programming languages. Thus the statements are more clear-cut and you can create them easier than the traditional programs.

The language is declarative which means both advantages and disadvantage. The disadvantage of the language for the students is the novelty and the singularity, because they learned procedural programming languages during previous studies. At the same time its advantage is that they did not have to write algorithm, they only have to write the query's result with help of the language and in the course of this procedure they have to think in sets not in steps ([13]).

The syntax of SQL is simple what can be deceiving because they think understood it. Nevertheless the learning of advanced SQL statements is much more difficult which can surprise them ([12]).

There are several problems can emerge during the course of learning SQL language. The problems described by [10] and [9] are the following:

- the students don't understand the relational model,
- the students don't remember well to the database schema and the names of tables and columns,
- the students have difficulties in making distinction between aggregate and scalar functions,
- the understanding of the concept of grouping and the observing the rules of grouping are difficult for the students,
- in join operations the writing of join conditions is difficult for the students.

Beside the not satisfactory understanding of the relational model and thinking in sets [7] note that the students don't understand the first order logic or rather some set operation causes problem for them.

In our opinion there are some more problems for students in the course of learning

- they haven't learnt yet the structure of SQL statements, reserved words and the order of clauses,
- they don't know well the syntax and semantics of SQL,
- they don't know well the relational operations,
- they don't get feedback; at the evaluation of a certain part of the SQL statements there are not results available (DDL, a part of DML, DCL), the students have to get used to that if they don't get error message in the course of running a statement, then the statement is syntactically correct, but it have to be

verified somehow if the statement results the desired operation,

- in the course of running a statement returned result the correctness of the result has to be verified; we mention joining of the tables as an example where the missing of a join operation generates more rows than number of real rows and between these rows there are some invalid relationship,

- they don't understand the error messages of database management system which problem occurs mainly at foreign language software ([4]).

We asked the students of ABK 2007 course in a questionnaire survey what caused them problems in the course of the learning of SQL. 46 of 56 students filled the questionnaire. It is random who did not answer, so we can consider the survey representative. We summarized in Table II the students' judgment about the phases of the learning of SQL ([3]).

Although SQL has a simple structure according to the data of Table II the understanding and observing the rules of SQL and coding SQL queries cause problems for one third of the students.

TABLE II.
THE PROPORTION OF THE STUDENTS WHO INDICATED PROBLEMS IN THE COURSE OF LEARNING OF SQL

Activity	Problem indications %
understanding language rules	30.1
observing and applying language rules	39.1
coding queries	43.5
writing WHERE conditions	8.7
using predicates	17.4
using aggregate functions	19.6
grouping	13.0
join operations	30.4
subqueries	39.1
creating and altering tables and other objects	10.9
updating data tables	8.7
data control statements	13.0

The highest ratio of join operation and subquery can be easily seen in Table II. Editing the statements with these elements is really not automatic, we have to think over which tables have to be joined and what will be the join conditions, or rather what we will query in subquery and in which clauses do we want to use the result of subquery. It is interesting that the students write down grouping fewer difficult though by our teacher experience the grouping and the selection of groups are problematic too for the major part of the students.

III. THE WAY TO SQL TUTORING SYSTEM AND THE ASSESSMENT OF THE SUPPORTING FORMS

Till the appearance of e-learning we had fewer tools to support the learning of SQL. The sample statements in SQL curriculum and ideal solutions saved in database management system as views help the individual learning. Students can compare their solutions with the ideal ones. The teacher consultation supports the group learning.

The students compare rarely their solutions with ideal solutions saved as views. This method of searching error and verifying the correctness of the statements is difficult, slow and requires strong concentration. It would support learning better if at the running of statements the students

got Hungarian feedback about correctness of statements and errors.

In our educational practice we use Moodle course management system from 2007 which enables new supporting forms as curricula, tests, tutorials and teacher feedbacks. We introduced these tools gradually into education hoping that these tools motivate the students better in the learning and make more efficient the acquirement of knowledge.

In 2007 in Moodle system we wrote 118 test questions helping learning of statement and rules of SQL language of which we compiled 7 tests grouping questions by topics.

We experienced that lower number of students solved these tests than data modeling tests. The reason behind is lack of time, because the students learn campaign-wise directly before exam paper ([5]). The writing of SQL exam paper is at the end of the semester and around this time there are other exam papers of other subjects, the deadline of submission of home works is in this period too, so the students have few time for solving tests.

On the Figure 1 we visualized the summarized number of solved tests of ABK 2007-2010 courses (column chart and left side axis) and the achieved results (line chart and right side axis). In examined period there were 81 students in this four courses, the number of solved tests is very low compared to the number of students. The number of test solved and the results both show that this supporting form of learning was not successful. For the efficient learning the students have to solve several tests more times.

At the ABK 2007 courses (three groups) we examined the results of exam papers and depicted the results on diagram, see Figure 2.

Figure 2 shows that while the major part of the students solved correctly the more simple queries there were
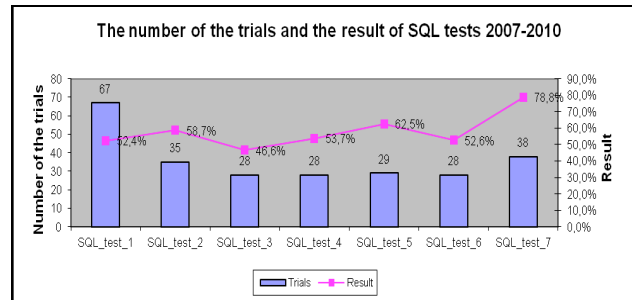

Figure 1. The number of the solved tests and the results at the ABK 2007-2010 courses
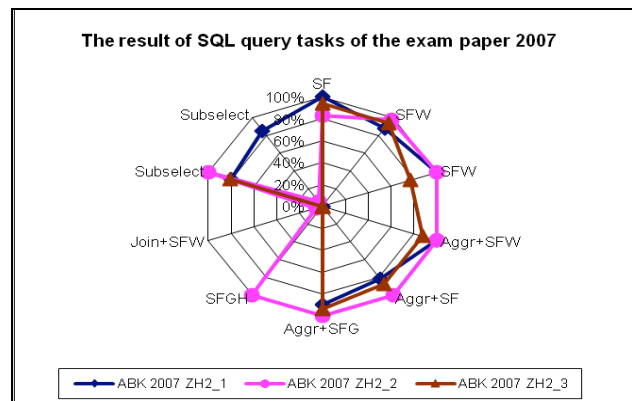

Figure 2. The results of SQL exam papers of ABK 2007 courses

queries such as grouping, join and subquery, which were not solved by any of the students in any group. This inspired us to write tutorials in topics of grouping, join, and subquery and set operation which explains the steps of editing of SQL statements in details. Control tests are also available for the tutorials too in which you have to solve query tasks in database management system and copy the statement into Moodle system. The teacher evaluates the statements but after closing tests we show the ideal solution as a feedback.

The students of ABK 2008 course learned with the help of the grouping, join and subquery tutorials, they opened a tutorial on average 3.8 times and one test 2.7 times. They solved only queries of grouping test at the other two tests they only viewed the ideal solutions feedbacks. In ABK 2009-2010 courses nobody used tutorials.

At ABK 2008 course we evaluated again the SQL exam papers. We experienced some improvement in grouping, join and subquery queries, but it could not be not proved that in this period the results were significantly better than the results of ABK 2007 courses ([6]).

We can't consider this supporting form successful either. That fact that the students don't require the teacher assessment of the queries, they only check the ideal solution feedback confirms that they would need feedback directly after solution of queries.

IV. THE SQL TUTORING SYSTEM

In the last 10-15 years many SQL tutoring system was wrote at various universities of the world. Some of them: AsseSQL ([1]), SQLator ([13]), SQLify ([7]), SQL-LTM ([8]), SQL-Tutor ([10]). These systems support SQL-92, give feedback for semantics of queries, evaluate the queries (heuristic, CQ), mark students and enable access SQL curricula.

We wanted to realize the following purposes with writing tutoring system by ourselves:

- the students can use it independently from place and time,
- help and signing of statement type can support the editing of the statement,
- curricula and tutorials can support the learning of SQL language,
- the students can get feedback about correctness of query,
- Hungarian error message can sign the occurrent errors.

The tutoring system is an ASP.NET application which was written in C# language, and which uses the ADO.NET object library to access MS SQL Server Express databases. The students used the tutoring system since the first semester of 2012/2013 school year.

The functions of the tutoring system:

- login and logout,
- database selection, showing of database schema and database description,
- question selection,
- editing and running SQL statement, displaying result, feedback about result, error message, displaying short and long help supporting editing statement or

rather clauses of ideal solution, selection of next question,
- displaying SQL curriculum,
- displaying SQL tutorials (grouping, join, subquery, set operations),
- sending e-mail to teacher.

The tutoring system works with SQL Server Express databases. At the present students can execute queries in three databases namely CD, School and Video. We store data needed to run the tutoring system as data of users, questions, ideal solutions, log, etc. in SQL Server Express database too.

At present there are 152 questions in database but we are going to write further questions too. The Table III shows the classification of the questions. Of course selection, grouping, aggregate function, ordering occur in other categories too.

TABLE III.
QUESTION CATEGORIES AND DISTRIBUTION OF QUESTIONS

Category	Operation	Number of questions	Distribution
I.	selection, aggregate function, ordering	60	39.5%
II.	grouping	27	17.8%
III.	join	26	17.1%
IV.	subquery	39	25.7%

In database we store the question in order of SELECT clauses and categories shown in Table III, we numbered them from one successively. The students can follow the questions in this sequence if they select the "next question" function.

In tutoring system now only SELECT statement can be run but any kind so you can edit and run statement step by step, but in this case the displaying of the correctness of the result will not be correct.

A short and long embedded help supports the edition of SQL statements in which we describe which relational operations have to be implemented. We assign a question type to every question which refers to clauses and language elements used in statement.

For example the SFWGJUA type means that beside SELECT, FROM, WHERE, GROUP BY clauses the ideal solution includes join operation, subquery and aggregate function too.

In case of errorless query or after three false running students can display the clauses of ideal solution. The aim of the restrictions is that the students can try statement editing and they do not begin their work with displaying of the clauses.

The tutoring system runs the student statement in SQL Server Express database management system than displays the result and finally compares it with the ideal solution. In the comparison it examines the fruition of $(H \setminus I) \cup (I \setminus H) = \varnothing$ connexion where H is the student solution and I the ideal solution. If the result of the left side operation series is the empty set and the element number of two result sets is equal than the result of two solutions is the same. This does not mean in every case that the solutions are equal too. For example you can realize a query of distinct rows with use DISTINCT key word or grouping too, or the result of some statement included

subquery can be equal with result of a statement included join.

V. HOW DO THE STUDENTS LEARN THE SQL LANGUAGE?

The SQL tutoring system logs every student operation we summarize how the students learn SQL language by this log table.

In autumn semester of 2010/2013 school year students of three courses used the SQL tutoring system: Computer Studies III. (KM3), Environmental Databases (KDB) and Applied Informatics (MM). The students decided themselves if they use the tutoring system or not so the students learned with tutoring system constitute random pattern.

In KDB course every student used the tutoring system so the result of their activity is representative. At the other two courses we examined with homogeneity test if the patterns have same distribution than the full courses considering the marks at the end of the semester. At both of courses we found that the distributions are same therefore the result of the activities is representative in this case too.

By data of Table IV in two courses (KM3, KDB) more than 90% of students learned with tutoring system while in the third course this proportion is 62%. This latter proportion can be explained as the first two courses are full time courses while the third is correspondence which students have more attended subject and have fewer time for learning beside their job.

Our prior expectation was that the students work for a longer time in a certain session namely 1-2 hours and they solve more tens of questions. However the data of log table show that the time of average session was 33 minutes and the students dealt with 9-12 questions on average considering data of all students.

Majority of the students solved questions of more databases. The questions of the databases are not the same difficult, you can find in CD database more simple, in Video more complicated and in School database the most difficult questions. Majority of the BSc students of KM3 course (77%) dealt with the questions of CD database, contra the students of other two courses solved rather the more difficult questions. We think that the MSc students of latter two courses learned in a more conscious way.

Although we taught SQL language for more than 20 years we never had information about how many running and error correction is needed to reach an errorless solution with good result after the editing of a SQL query.

TABLE IV.
QUESTION SELECTION AND RUNNING DATA BY COURSES

	KM3	KDB	MM
number of students in the course	16	11	21
used the tutoring system	15 (94%)	11 (100%)	13 (61.9%)
total number of login	99	99	137
total number of question selection	913	1172	1711
question selection per login	9	12	12
total number of running	1814	2790	3925
syntactically errorless	929 (51.2%)	1464 (52.5%)	2259 (57.6%)
errorless and good result	455	841	992

		(25.1%)	(30.1%)	(25.3%)
number of students who ran queries more than 30 times		6 (40%)	10 (90.9%)	9 (69.2%)
average running per students (>30)		280	279	430
average running per students (<30)		15	5	14

In Table IV data of all three courses show that near half of the total running is syntactically incorrect and almost one third of total running is errorless and has good result. If we compare the latter data with the number of question selection we see that in case of almost 30% of the questions students give up and do not get to the correct solution. This 30% includes those cases too when the result is not comparable with the ideal solution because in set operation the number or type of elements of SELECT lists are not the same.

In each course there were some students who dealt with only a relatively low number of questions. We determined separately the average number of runnings among those who run fewer than 30 times and who run more than 30 times as reflects better the reality than the course average. The number of runnings of those students who run more than 30 times shows that these students were motivated in learning while students who run less than 30 times not.

There were students in every course who solved all queries some of them was solved more times. The maximum number of running and the maximum number of errorless solution with good result are the following: 479 – 136 (KM3), 530 – 191 (KDB) and 749 – 188 (MM). All of these students wrote flawless exam paper which confirms that the SQL tutoring system is useful for learning SQL language.

TABLE V.
DISPLAYING DATABASE DESCRIPTION AND DIAGRAM

Operation	KM3	KDB	MM
login	99	99	137
displaying database description	16 (16.1%)	20 (20.2%)	57 (41.4%)
displaying database diagram	33 (33.3%)	58 (58.6%)	23 (16.8%)

We summarized in Table V that students how many times opened the database description and database diagram after login. The databases are simple with 4-6 tables at denomination we tried to select talkative names so it is enough to read the description only once. Based on the data of the table and considering the number of students the students seems to do it.

At the teaching of SQL language in the course of class we began to work with displaying database diagram which show the names of tables and columns, the data types of columns and the relationships between tables helping the editing of queries. As the data of the table show in the course of their individual work the students did not do so. This can explain that the most frequent error is the invalid column name (see Table IX).

In Table VI we classified the question selections by question category – Table III shows the question categories. The data shows that the students begin editing of queries with first questions and they solve most

questions from first category. Only a minor part of the students get to the questions of other categories.

TABLE VI.
QUESTION SELECTION BY QUESTION CATEGORY

Category	KM3	KDB	MM
I.	55.0%	47.5%	58.0%
II.	20.5%	17.3%	18.4%
III.	11.6%	16.2%	12.4%
IV.	12.9%	18.9%	11.3%

Table VII shows which helping tools students used in the course of query editing. We expected that the number of displaying short and long helps will clearly exceed the number of displaying clauses especially considering that the displaying of the latter one is restricted as we mentioned it above. In contrast with it in two courses less than ¼ of the students checked the short help and less than 1/5 of the students checked the long help. In the third course the number of students who checked the helps was negligible. In each three course the number of checking clauses was 60% or above.

TABLE VII.
DISPLAYING HELPS AND CLAUSES

Operation	KM3	KDB	MM
question selection	913	1172	1711
displaying short help	222 (24.3%)	254 (21.7%)	63 (3.7%)
displaying long help	159 (17.4%)	203 (17.3%)	65 (3.8%)
displaying clauses	547 (59.9%)	778 (66.4%)	1481 (86.6%)

Table VIII shows how many views of clauses happened in the certain courses by question category. If we have a look on the highest values then every area of SQL learning was problematic for students of some courses:

- writing of selection condition of WHERE clause and using aggregate function for students of KM3 and MM courses,
- grouping for students of MM course,
- join for students of KM3 and KDB courses,
- use of subqueries for students of KDB course.

TABLE VIII.
CLAUSE SELECTION BY QUESTION CATEGORIES

Category	KM3	KDB	MM
I.	183 (33.5%)	144 (18.5%)	457 (30.7%)
II.	99 (18.1%)	184 (23.7%)	428 (28.7%)
III.	170 (31.1%	196 (25.2%)	331 (22.2%)
IV.	95 (17.4%)	254 (32.6%)	275 (18.4%)

In case of every three courses we examined that which clauses were checked by students most times at certain categories. We experienced that at the first category (selection condition and aggregate functions) they checked the WHERE clause most of the time. At the second category (grouping) they checked the GROUP BY clause most often, at the third category (join) FROM clause while at the fourth category (subquery) WHERE clause which includes subqueries. These are in line with our expectations. However we did not expect that the second most displayed clause was SELECT clause. When we asked students' opinion about tutoring system they

indicated that was not obvious at every question what columns had to appear in SELECT clause. There was frequent error at the verifying of the result that the SELECT column lists were not same at student' solution and ideal solution which due to correction purposes gave motivation to check the SELECT clause too.

The other types of help were used only in a negligible number of cases during the 335 login in total. In the three courses together they opened SQL curriculum only 16 times, they viewed tutorials 31 times and not more than 2 times they sent e-mail. All of tutorials were opened once or twice, but they mostly checked subquery tutorial.

In Table IX we summarized the most frequent running errors. The most frequent two errors are syntax error which refers to not satisfactory knowledge of syntactical rules of SQL statements and invalid column name which indicates that the students do not remember the data model properly. The other errors also imply that students do not know satisfactorily the rules of SQL language. We think that the number of errors could be decreased substantially if students used available description and curricula more often.

TABLE IX.
FREQUENCY OF RUNNING ERRORS

Description of errors	KM3	KDB	MM
syntax error	24.40%	24.1%	32.8%
incorrect keyword	12.70%	17.6%	11.8%
invalid column name	36.60%	22.8%	18.8%
ambiguous column name	5.60%	5.7%	2.5%
condition without logical expression	6.60%	6.1%	9.8%
the column does not occur in aggregate function's argument or GROUP BY clause	4.30%	8.9%	11.7%

By the data of Table X the result corresponds to the result of the ideal solution only at almost half of the running of syntactically errorless statements. The result cannot be verified if the statement includes ORDER BY clause or if its SELECT list is not the same as the SELECT list of ideal solution. Syntax and conversion error occurred in very low percentage which refers to program error what we are going to repair.

TABLE X.
VERIFYING RESULT

The result of verifying	KM3	KDB	MM
cannot be verified because of the ordering	14.4%	5.8%	5.4%
In SELECT lists the number of columns or type of columns are not equivalent	10.9%	11.5%	10.9%
syntax or conversion error	0.2%	0.2%	0.5%
the results are not the same	25.5%	25.1%	39.2%
the results are the same	49%	57.4%	43.9%

VI.   CONCLUSION

It proved to be a good practise to introduce tutoring system into education. We executed three independence tests where we compared the results of exam papers of three courses with results of exam papers of courses of previous year since the curricula, the teacher and the exam

paper were the same in the two years. The only difference was the use of tutoring system ([11]).

TABLE XI.
THE RESULTS OF THE INDEPENDENCE TESTS

	KM3	KDB	MM
calculated probability	0.005533	0.014166	0.264859
χ^2	18.29626	20.67661	12.30978
degree of freedom	6	9	10
significance level	0.05	0.05	0.05
critical value	12.59159	16.91898	18.30704

By the data of Table XI in case of two courses (KM3, KDB) the result is significantly better, at the third course (MM) we could not prove that.

The students worked with tutoring system gladly and effectively. We asked the students to evaluate the tutoring system on a five degree scale. The students of KM3 courses evaluated tutoring system 4.75, the students of KDB courses evaluated it 4.85 while this value was 4.2 at the students of MM courses.

The students gave textual assessment too:

- The program is very simple to handle, transparent and really well useable.
- With the program you can prepare for writing exam paper effectively.
- The program offers serious help to understand the SQL language.
- With using program the SQL language turned out to be more transparent.
- It would be great to do the exam paper in the tutoring system.

Beside the positive opinions above students noted that it was disturbing for them that they had to write exam paper in another database management system namely in Access although we drew attention to differences between SQL dialects of two database management system. The other error was that the running of program failed from time to time due to a program error. It was our first multiuser web application development and after a time you could test the program only in real environment which explains the errors. Of course we repaired and repair the program errors continuously.

We plan to do further corrections and expansions:

- Before comparing the results we will delete the ORDER BY clauses in order to set operations to be doable.
- By the same purpose we will verify the SELECT list with the program and we will change the order if necessary.
- We will examine and if possible we will realize the possibility the writing of exam paper.
- We will review the questions in order to make it obvious what should appear in SELECT lists.
- We plan to make new databases and new questions.

REFERENCES

[1] Coleman Prior, J.: Online Assessment of SQL Query Formulation Skill, Proceedings of the fifth Australasian conference on Computing education, 2003, Vol. 20, 247-256

[2] Czenky, M.: Adatmodellezés, SQL és Access alkalmazás, SQL Server és ADO [Data modeling, application of SQL and Access, SQL Server and ADO], ComputerBooks Kiadó, Budapest, 2005, p. 445

[3] Czenky, M.. Adatbázis-kezelés oktatás hallgatói szemmel [The Education of the Database Management with Student Eye]. Proceedings of Informatika a felsőoktatásban 2008 Konferencia, Debrecen, Hungary. p. 9. Retrieved January 27, 2012 from http://www.agr.unideb.hu/if2008/kiadvany/papers/A45.pdf

[4] Czenky, M.: Az SQL nyelv tanítása [Teaching of SQL Language], Proceeding of I. Informatika Szakmódszertani Konferencia, Szombathely, 2008, CD, p. 13

[5] Czenky, M.: Hallgatói tevékenységek elemzése a Moodle naplófájljai alapján [The analysis of students activities based on Moodle log files], MoodleMoot 2010 Konferencia, Gödöllő

[6] Czenky, M.: The Examination of the Efficiency of the Teaching of SQL, Kitekintés - Perspective, XV. évfolyam, különszám, Békéscsaba, 2011, pp.244-252

[7] Dekeyser, S., de Raadt, M. and Lee, T. Y.: Computer assisted assessment of SQL query skills, ACM International Conference Proceeding Series; Vol. 242, Proceedings of the eighteenth conference on Australasian database - Volume 63, Ballarat, Victoria, Australia, 2007, p. 53 – 62

[8] Dollinger, R.: SQL Lightweight Tutoring Module – Semantic Analysis of SQL Queries based on XML Representation and LINQ, World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2010, pp. 3323-3328

[9] Kearns, R., Shead, S., and Fekete, A.: A teaching system for SQL. In Proceedings of the 2nd Australasian Conference on Computer Science Education. (Melbourne, Australia, 2 - 4 July 1997), p. 224 - 231.

[10] Mitrovic, A.: A Knowledge-Based Teaching System for SQL, EdMedia World Conference on Educational Media and Technology 1998. Retrieved February 10, 2013 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1046

[11] Pestman, W. R.: Mathematical statistics: an introduction. Berlin, New York: de Gruyter, 1998, p.545

[12] Renaud, K. and van Biljon, J.: Teaching SQL —Which Pedagogical Horse for this Course? In proceeding of Key Technologies for Data Management, 21st British National Conference on Databases, BNCOD 21, Edinburgh, UK, July 7-9, 2004, p. 244-253

[13] Sadiq, S., Orlowska, M., Sadiq, W. and Lin, J.: SQLator – An Online SQL Learning Workbench, ACM SIGCSE Bulletin, Volume 36, Issue 3, 2004, p. 223 – 227

[14] J. D. Ullmann and J. Widom, Adatbázisrendszerek – Alapvetés [Database systems - basics]. Budapest: Panem KFT, 2008.

AUTHOR

M. Czenky is an assistant professor at the Department of Informatics, Szent István University, Gödöllő, Hungary, 2100 (e-mail: marta.czenky@t-online.hu).